QQI

Higher Diploma in Science in Data Analytics

AUTUMN 2019 EXAMINATIONS

Module Code: **B8IT108**

Module Description: **Data and Web Mining**

Examiner: **Terri Hoare**

Internal Moderator: **Dr Shahram Azizi Sazi**

External Examiner: **Dr Catherine Mulwa**

Date: Thursday, 31st October 2019
Time: 18:30-20:30

INSTRUCTIONS TO CANDIDATES

Time allowed is 2 hours
Question 1 is mandatory (30 marks)
Answer 2 of the remaining 3 Questions (35 marks each)

All answers should reference literature and case studies as appropriate. Use of scientific calculators only is permitted.

Question 1: MANDATORY **(30 Marks)**

a) Use the AllElectronics Customer Database Table below to build a Naïve Bayes Classifier that can be applied to predict the class label buys_computer for the unlabelled tuple X given below.

X = (age=youth, income=medium, student=yes, credit_rating=fair)

Hint: apply the Naïve Bayesian formula

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

$$where \;\; P(X|C_i) = P(x_1|C_i) * P(x_2|C_i) * \ldots * P(x_n|C_i)$$

C_1 *buys computer*
C_2 *does not buy computer*

Electronics Customer DB - Training Tuples - Label buys_computer

RID	age	income	student	credit_rating	buys_computer
1	youth	high	No	fair	**NO**
2	youth	high	No	excellent	**NO**
3	mid-aged	high	No	fair	**YES**
4	senior	medium	No	fair	**YES**
5	senior	low	Yes	fair	**YES**
6	senior	low	Yes	excellent	**YES**
7	middle-aged	low	Yes	excellent	**YES**
8	youth	medium	No	fair	**NO**
9	youth	low	Yes	fair	**YES**
10	senior	medium	Yes	fair	**YES**
11	youth	medium	Yes	excellent	**YES**
12	middle-aged	medium	No	excellent	**YES**
13	middle-aged	high	Yes	fair	**YES**
14	senior	medium	No	excellent	**NO**

[10]

b) Decision trees are one of the most intuitive and frequently used predictive algorithms in practice. A sample dataset has four attributes (temperature; humidity; wind; and outlook) that are used to predict whether to play golf (yes or no). Information Gain for all attributes has been calculated and is presented in the table below. Show the Information Gain calculations for **Outlook**, the root node of the decision tree.

Hint:-

$$Information\ Gain(A) = Info(D) - Info_A(D)$$

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

Attribute	**Information Gain**
Temperature	0.029
Humidity	0.102
Wind	0.048
Outlook	0.247

Outlook	Temperature	Humidity	Wind	Play Golf
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	78	false	yes
rain	70	96	false	yes
rain	68	80	false	yes
rain	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rain	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rain	71	80	true	no

[10]

c) The objective of a model is to predict whether a person is likely to respond to a direct mailing campaign or not based on demographic attributes (age, lifestyle, earnings, type of car, family status, and sports affinity). The validation dataset is labelled either "response" or "no response". The confusion matrix for the validation dataset is given below. Calculate the related performance measures of accuracy, recall, precision, specificity, and sensitivity.

	Actual no response	Actual response
Predicted no response	1231	146
Predicted response	394	629

[10]

ANSWER (2 OUT OF 3) QUESTIONS BELOW

Question 2 **(35 Marks)**

a) Select and critique an appropriate algorithm for building a model using supervised learning. Your answer should include a description of how the algorithm works, input and output limitations, pros, cons, hyper-parameters, and types of suitable applications.

[20]

b) Describe and critique cross validation as the gold standard for evaluating the training performance of a model.

[7]

c) Ensemble models improve the error rate and reduce the bias of individual models by aggregating (voting between) the predictions of several base models to produce the final model.

Propose four ways to ensure base models are diverse.

[8]

Question 3 **(35 Marks)**

a) Select and critique an appropriate algorithm for building a model using un-supervised learning. Your answer should include a description of how the algorithm works, input and output limitations, pros, cons, hyper-parameters, and types of suitable applications.

[20]

b) Discuss normalisation in the context of clustering algorithms and propose and describe a method for normalising data.

[6]

c) Propose a use case for a SOM (Self Organising Map) in data mining.

[5]

d) Discuss the role of support and confidence in mining association rules.

[4]

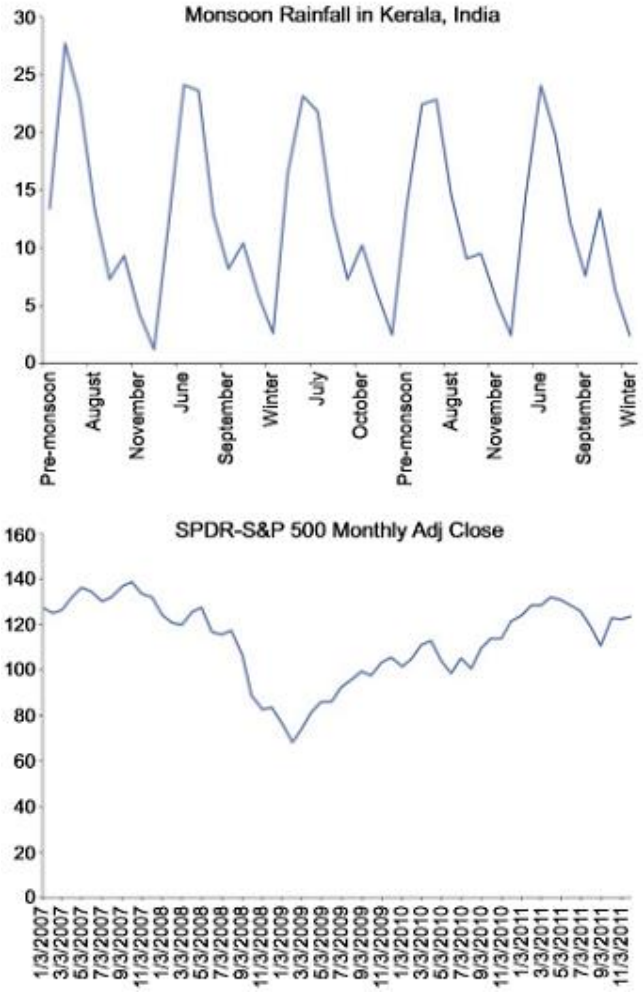
Question 4 **(35 Marks)**

a) PageRank is an algorithm used by Google Search to rank websites. Critique the measure of network centrality that it uses and contrast with at least two other measures of 'centrality' or power in a network graph.

[15]

b) The figures below show two different time series: annual monsoon precipitation in Southwest India averaged over a five-year period and the adjusted month-end closing prices of the SPDR S&P 500 (SPY) Index over another five-year period. For each time series, propose either a data or model driven forecasting approach motivating the selection of the approach in each case.

[8]



c) Discuss the role of TF-IDF in text mining.

[6]

d) Propose a supervised learning approach for performing sentiment analysis of blogs.

[6]

END OF EXAMINATION